Comparison of Several Different Non-Disclosure Methods for Tabular Data Using a "Real Life" Table Structure of Complex Hierarchies and Links

Ramesh A Dandekar
EI-70, U. S. Department of Energy, Washington DC 20585[1]
Ramesh.dandekar@eia.doe.gov

Abstract

To allow comparison of various sensitive tabular data protection methods on a consistent basis, the statistical disclosure control/limitation (SDC/SDL) researchers have long used public domain artificial (synthetic) data sets available from http://webpages.ull.es/users/casc/ website. The format used by these data sets, however, fails to convey visualization aspects of inherent complexities associated with various structural details typical of public use tables. The practitioners of tabular data protection methods in federal statistical agencies have some familiarity with commonly used table structures. However, they require some guidance on how to evaluate appropriateness of various sensitive tabular data methods when applied to their own table structure. With that in mind, we use a real life "typical" table structure of moderate hierarchical and linked complexity and populate it with synthetic micro data to evaluate the relative performance of four different tabular data protection methods. The methods selected for the evaluation are: 1) lp-based classical cell suppression 2) lp-based CTA (Dandekar 2001), 3) network flow-based cell suppression as implemented in DiAna, a software product made available to other Federal statistical agencies by the US Census Bureau and 4) a micro data level noise addition method documented in a US Census Bureau research paper (Evans, Zayatz, and Slanta 1998). The outcome from the comparative evaluation is available from http://mysite.verizon.net/vze7w8vk/

The classical lp-based cell suppression method used for the evaluation is similar to that used by CONFID at Statistics Canada since the mid-80. The selection of the complementary cell suppression pattern is done by using a cost proportional to the table cell value as an objective function. This results in higher preference for smaller tabular cells as complementary suppression cells.

The controlled tabular adjustments (CTA) a.k.a. synthetic tabular data method used is the one documented in Dandekar (2001) and Dandekar/Cox (2002) . Large size non-sensitive table cells are targeted for adjustments by using a cost function which is a reciprocal of the table cell value. Such an approach results in relatively small percentage changes in the cell values and therefore, reduces the overall degradation in the accuracy of the statistical information imbedded in table cell values.

The network flow model in the DiAna software uses a minimal cost flow (mcf) based algorithm from the University of Texas to develop a complementary cell suppression pattern. The PC version of the software used for this evaluation targets smaller sized cells to develop a complementary cell suppression pattern.

The micro data level noise addition method as described in (Evans, Zayatz, and Slanta 1998) is used for this evaluation. Micro data is perturbed by an average of 10% and standard deviation of 0.005 by using a normal distribution.

Questions For the Committee

1. As users of data, would you prefer to have tables 1) protected by methods that change the data slightly, such as CTA or noise adjustment, or 2) protected by suppression – withholding sensitive cells plus others to protect them?
2. What are the issues that EIA should consider as we try to come up with a common approach to protecting tabular data?
3. If EIA wants to begin using data adjustment methods (CTA or noise), how should EIA inform users that the data have been changed?

[1] "*This is a working document prepared by the Energy Information Administration (EIA) in order to solicit advice and comment on statistical matters from the American Statistical Association Committee on Energy Statistics. This topic will be discussed at EIA's fall 2007 meeting with the Committee to be held October 18 and 19, 2007."

References

Dandekar R. A. (2001) "Synthetic Tabular Data: A Better Alternative To Complementary Data Suppression" - Original Manuscript. Energy Information Administration, U. S. Department of Energy. Also available from CENEX-SDC Project International Conference, PSD2006, Rome, Italy, December 13-15, 2006, Companion CD Proceedings ISBN: 84-690-2100-1.

Dandekar R. A. and Cox L. H. (2002), Synthetic Tabular Data: An Alternative to Complementary Cell Suppression, 2002. Manuscript, Energy Information Administration, U. S. Department of Energy.

Dandekar, R.A (2003), Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems, working paper 40, UNECE Work session on statistical data confidentiality (Luxembourg, 7-9 April 2003)

Dandekar Ramesh A. (2004), Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data, pp 121-135, Lecture Notes in Computer Science, Publisher: Springer-Verlag Heidelberg, ISSN: 0302-9743, Volume 3050 / 2004**,** Title: Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004, Barcelona, Spain, June 9-11, 2004.

Evans T., Zayatz L., Slanta J. (1998),"Using Noise for Disclosure Limitation of Establishment Data", Journal of Official Statistics 1998. Also available from http://www.census.gov/srd/papers/pdf/bte9601.pdf

Fischetti, M. and J. J. Salazar (2000), "Models and Algorithms for Optimizing Cell Suppression Problem in Tabular Data with Linear Constraints", *Journal of the American Statistical Association* **95**, 916-928.